WHAT IS CLAIMED IS:

1. A method for operating a server cluster comprising N server nodes to service client requests, each client request being directed to one of a plurality of sites hosted on said server cluster, each site being identified by a domain name and each server node being identified by an address on a network connecting said clients to said server nodes, said method comprising the steps of:

measuring the computational resources required to service said requests to each of said sites over a first time period;

grouping said sites into N groups, each group being assigned to a corresponding one of said server nodes such that for each pair of groups, the difference in the sum of said measured computational resources is within a first predetermined error value; and

providing configuration information to a router accessible from said network, said information defining a correspondence between each of said sites and one of said server nodes assigned to one of said groups containing that site, said router providing said address of said server node in response to a message specifying said domain name of said site.

2. The method of Claim 1 wherein said router is a Domain Name System (DNS) server.

3. The method of Claim 1 wherein said sites return files in response to said requests, and wherein said step of measuring said computational resources comprises recording information identifying each returned file, the size of that file, and the number of times that file was returned.

4. The method of Claim 3 wherein each of said server nodes comprises a cache memory for facilitating the return of said files in response to said request and wherein said step of grouping said sites also depends on the amount of memory in said cache memory on each of said servers.

5. The method of Claim 4 wherein said groups are chosen such that said files returned during said first time period more than a predetermined number of times can be stored simultaneously in said cache memory.

6. The method of Claim 3 wherein said measurement of said computational resources further comprises measuring the number of bytes of data returned in response to said requests for each site during said first time period.

7. The method of Claim 6 further comprising estimating the number of bytes of data returned directly from said cache memory in servicing said requests for each site during said first time period.

8. The method of Claim 1 wherein one of said sites belongs to two of said groups.

9. The method of Claim 1 wherein one of said sites belongs to all of said groups.

10. The method of Claim 7 wherein said router selects which of said service nodes corresponding to said two groups will service a request for that site.

11. The method of Claim 1 further comprising the steps of:

measuring the computational resources required to service said requests to each of said sites over a second time period; and

grouping said sites into N new groups, by swapping sites between said previous groups, each new group being assigned to a corresponding one of said server nodes such that for each pair of new groups, the difference in the sum of said measured computational resources over said second time period is within a second predetermined error value.

12. The method of Claim 11 wherein said new groups differ from said previous groups by as few site swaps as possible.